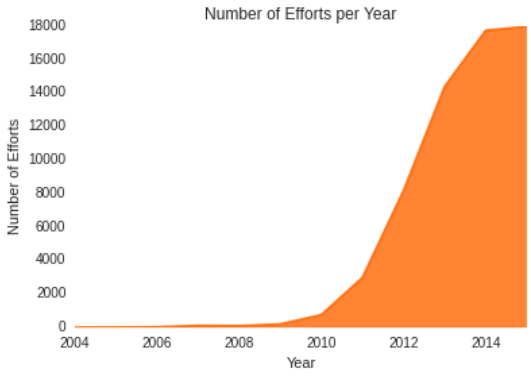
Predicting Bicycle Speeds on Mission Bay with Strava

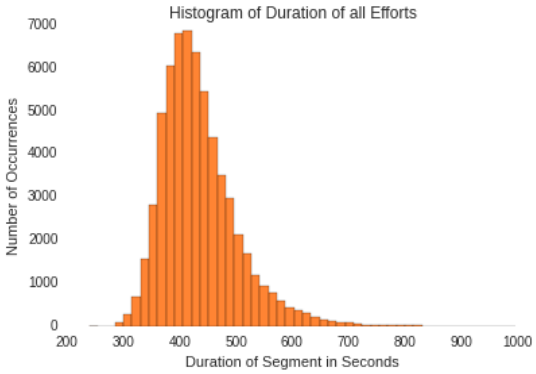
Rachel Marty
A09162716
CSE 255

ABSTRACT
Strava is a popular GPS-based cycling application. Using data gathered from this application, I predict the duration of time used by particular riders at specific times to traverse a 2-mile route along Mission Bay in San Diego.

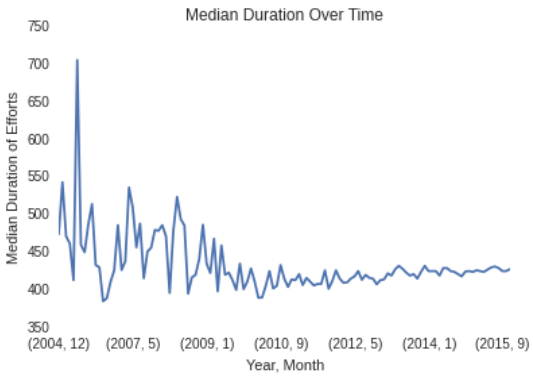
DATASET
The dataset used for this project was extracted from the Strava API. Strava, a popular GPS-tracking application for cyclists, captures rides by focusing on "segments" – a path commonly travelled by several riders. The data in this project focuses on a segment along Mission Bay in San Diego. This segment is about 1.7 miles long and is popular among commuters, competitive athletes, and leisurely riders. Since 2004 when Strava was established, over 60,000 rides have been taken from more than 9,000 people. As the application has become more popular, the number of rides has been increasing exponentially.



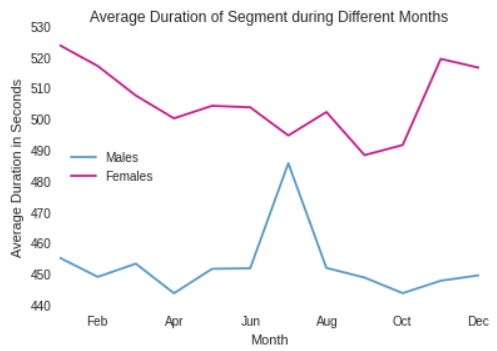
The distribution of durations for each effort is parametric with a long right tail.



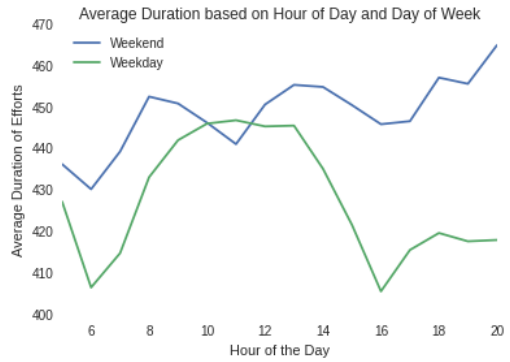
While the number of efforts has been increasing, the average duration of each effort has also been changing over time. As the seasons change, there are shifts in speeds. However, as the number of efforts increases, more stability in duration of time taken is achieved.



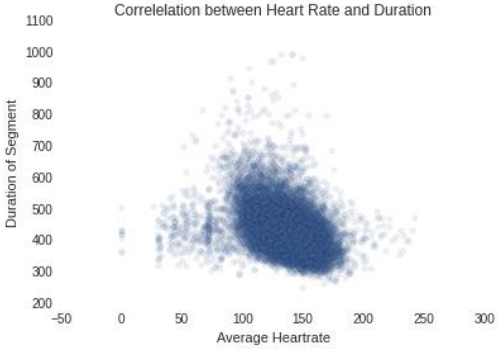
Universal temporal data plays a significant role in the average speeds of cyclists on Mission Bay. Although San Diego's climate is not highly variable, the month of the year has impact on speed. Interestingly, the changing seasons and presumably weather affect males and females differently. While men are consistently faster than women, the trends within the gender groups vary inversely.



Furthermore, the day of the week impacts the speed of riders at different times of the day. Looking at the time of the day can further stratify these speeds. While the middle of the day as similar segment durations, the cyclists travel significantly faster during the morning and evening on weekdays than on weekends. This observation can likely be attributed to the difference between commuters and leisurely riders.



Average heart rate during a ride also has an inverse correlation to the duration of the effort with a Pearson R value of -0.39. Cyclists who push their heart rates higher tend to be riding faster.



Finally, cyclists can annotate their ride with a title. Some of the words they choose to include in their titles correlate with faster or slower rides.

PREDICTIVE TASK

Since there appears to be several features that are indicative of speed, I will work on the problem of predicting the amount of time a cyclist takes to ride the Mission Bay segment on a particular attempt. As discussed in the Literature section, there are two popular ways for cyclists to estimate the amount of time for a route – Google Maps and Strava route prediction. Strava uses a simple prediction strategy, but I cannot access the necessary data to replicate it, so I am unable to compare to their method. However, I will use the features claimed by Google Maps for their time duration prediction. Furthermore, I will compare my method to the simple baseline of the average of duration of all riders and to the more personalized method of the average duration of individual riders.

I compared all of the methods using the mean absolute error due to some large outliers in the dataset. Using five fold cross validation, I constructed a training set from 80% of my data and a test set from 20% of my data. After training and testing on each five sets, I averaged the mean absolute errors from all to get an average training error and an average testing error. Then, I selected the regularization parameter based on a minimization of the mean absolute error of test set.

As described in the Model section, I chose to employ a regression with features about the individual, previous rides and the time and date

of the ride. I chose a regression because I am predicting continuous output. The Strava API provides access to a limited amount of information about each effort. In order to collect sufficient data to train my model, I had to also download information about all cyclists and the rest of each cyclist's rides. I merged all of this information in order to create my features.

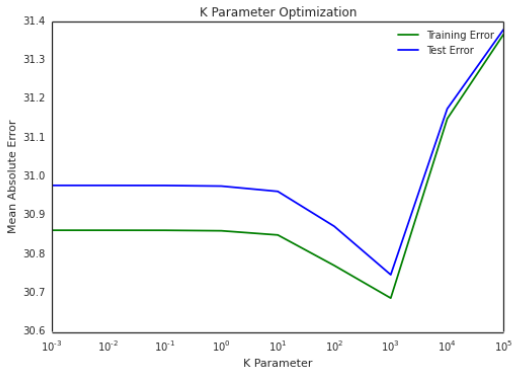
MODEL

I chose to use a Ridge Regression model.

$$Ø = (X^TX + kI)X^TY$$

I compared the Ridge Regression model to a simple Linear Model. However, the Ridge Regression outperformed the simple Linear Model by a slight margin. Instead of using a complex model, I carefully selected features in order to optimize performance. The features will be discussed in detail in the results section.

Once the features were chosen, I trained the Ridge Regression model using several values of k and selected the model with the lowest average mean absolute error on the test set. Originally, I had been using k = 1; however, I was over-fitting the model to the training data. For the chosen features, the optimized value of k was 1000.



Originally, I made the assumption that cyclists would improve over time. In order to capture this expertise, I planned to use a sliding window. However, the actual data did not reflect this hypothesis – cyclists had varied segment durations but did not improve significantly over time. Even using the immediate surrounding durations by the cyclists proved to have less predictive power than their overall average.

LITERATURE

Predicting travel time for cars in traffic is a well-studied problem. Google Maps is known for gathering speed limits, recommended speeds, road types, historical average speeds, actual travel time from previous users, and traffic information to accurately predict travel time in a car[1]. However, training models to predict cycling speed on a given route presents a new set of challenges. Travel time for cars is highly dependent on traffic and independent of the car itself. On the other hand, travel time on a bicycle is less dependent on traffic and more dependent on the features of the individual rider.

The two most popular venues for predicting travel time on a bicycle are Google Maps and Strava. However, both methods are naïve and do not exploit all available data. Google Maps employs a strategy that does not consider the individual, rendering the predictions very inaccurate[2]. First, they start with a baseline travel speed. Then, they modify the baseline speed slightly to account for inclines and declines on the route. Finally, they add some time to the prediction to account for stop sign and stop light delays. Strava, a popular cycling and running GPS platform, only uses information about the individual. Strava calculates your average speed over the prior four weeks and estimates your travel time of the new route to be at your average speed[3]. It does not account for traffic or incline because it assumes those factors will also be at play in your prior rides.

Very little research seems to have been done to predict cycling travel times with more specificity than these two methods. Instead, much of the focus on cycling data has been on predicting the possible output of the most competitive athletes. However, research has

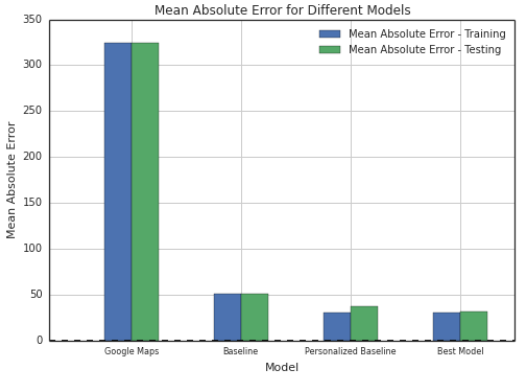
been done to predict running pace from individualized data[4]. The algorithm includes the following features: elevation gain, run mileage, and most recent 10k time. Several regression models were compared and basic linear regression performed the best on the test set.

My conclusions are very different than either the Google Maps or Strava solution because I employ a vastly different feature set. Predicting travel time on a bicycle is likely not a priority for either of these companies because they neglect the sophistication that would vastly improve their results.

RESULTS

As discussed above, I am comparing my best model to three other models. First, I compare it to my implementation of the Google Maps method. Google Maps takes a very simple approach of cyclists. They assume an average speed of 16km/hour for every rider. Since the chosen segment has no incline and no stoplights, the method does not deviate from this average. However, this pace is very slow as compared to the cyclists using Strava. Thus, the Google Maps prediction does very poorly as measured by the mean absolute error. Second, I constructed a simple baseline for comparison. I predicted the mean duration across all efforts on the segment. While this prediction significantly outperforms the Google Maps prediction, it does not include any personalization. Third, I constructed a personalized baseline for comparison. While I could not duplicate the Strava prediction methods due to insufficient available data, this method comes the closest to it. For each cyclists, I predicted the average of their prior attempts. Using very little data, this method performs relatively well. While the error is as low as the best method, Strava probably uses a similar method due to the low overhead cost.

The Best Model uses a complex feature vector trained with Ridge Regression. First, the feature with the most weight is the average duration of every effort made by each individual cyclist. Then, since temporal values highly influence duration of the effort, I constructed features for each hour of the day, day of the week, and month of the year. Each effort is assigned zeros for all of these features except for the specific time when the ride occurred. Next, I created features that reflect attributes of the individual riding. These attributes include the cyclist's gender (since males in general ride faster than females), their average spinning cadence, their max overall speed, average heart rate, and whether or not they pay for a premium Strava membership. Cyclists who spin faster and maintain a higher heart rate tend to spend a shorter duration on the segment. Furthermore, users who invest in a premium membership tend to ride faster because they are more dedicated to the sport. Finally, after each ride, the cyclist writes a description. The words chosen in this description have good predictive power – alone the features based on the word choice can beat the baseline predictor. Certain words like "Soledad" and "Torrey" are associated with faster efforts, presumably because these are difficult hills ridden only by experienced riders. Other words, like "today", are used more commonly by significantly slower riders.



Model	Mean Absolute Error - Training	Mean Absolute Error - Testing
Google Maps	324.393972	324.393974
Baseline	51.057718	51.057714
Personalized Baseline	30.269281	37.460748
Best Model	30.770846	30.871741

Based on the dataset exploration, I assumed that the temporal data would be the most powerful predictor. However, once the

individual averages were added to the model, the temporal data added relatively little. Instead of time dictating speed, faster riders simply tend to cycle at different times than slower riders. Furthermore, adding a gender feature added relatively little since there are few female riders and most of them ride frequently enough to have accurate averages. Surprisingly, the user inputted descriptions were also very predictive – outperforming the baseline alone.

Overall, the Best Method outperforms Google Maps and both baselines due to the features it employs. The combination of personal averages and personal attributes in a regression predicts the duration of any specific effort with high accuracy.

REFERENCES

1. Russell, Richards. "How Does Google Maps Calculate Your ETA?". Forbes, Tech. http://www.forbes.com/sites/quora/2013/07/31/how-does-google-maps-calculate-your-eta/
2. Lobo, Adrian. "How accurate are Google Maps cycling time estimates?" Better by Bicycle. http://www.betterbybicycle.com/2014/09/how-accurate-are-google-maps-cycling.html
3. Anderson, Elle. "Question about Estimated Moving Time for a Strava Route". Strava, Help Center. https://strava.zendesk.com/entries/28103400-Question-about-Estimated-Moving-Time-for-a-Strava-Route-
4. Jin, Tiffany. "Predicting Pace Based on Previous Training Runs". http://cs229.stanford.edu/proj2014/Tiffany%20Jin,%20Predicting%20Pace%20Based%20on%20Previous%20Training%20Runs.pdf